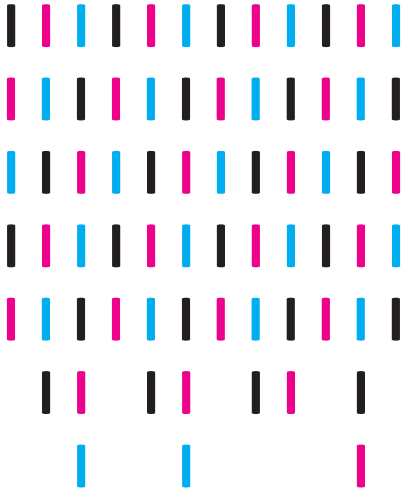

NEPS SURVEY PAPERS

Timo Gnambs

NEPS TECHNICAL REPORT FOR ENGLISH READING COMPETENCE: SCALING RESULTS OF STARTING COHORT 3 FOR GRADE 10

NEPS Survey Paper No. 31
Bamberg, November 2017

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

Survey Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Survey Paper Series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and NEPS.

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS Survey Papers are available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

NEPS Technical Report for English Reading Competence: Scaling Results of Starting Cohort 3 for Grade 10

Timo Gnambs

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T. (2017). *NEPS Technical Report for English Reading Competence: Scaling Results of Starting Cohort 3 for Grade 10* (NEPS Survey Paper No. 31). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Acknowledgements:

Various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced *verbatim* from previous working papers (e.g., Haberkorn et al., 2012; Hardt, K., Pohl, S., Haberkorn, K., & Wiegand, E., 2013) to facilitate the understanding of the presented results.

NEPS Technical Report for English Reading Competence: Scaling Results of Starting Cohort 3 for Grade 10

Abstract

The National Educational Panel Study (NEPS) investigates the development of competencies across the life span. Therefore, the NEPS develops tests for the assessment of various competence domains in different age cohorts. In order to evaluate the quality of these competence tests, several analyses based on item response theory (IRT) are performed. This paper describes the data and scaling procedures for a reading competence test on English as a foreign language in grade 10 of starting cohort 3 (fifth grade). The reading competence test included 13 items (distributed among three booklets) with multiple choice response formats and matching tasks that represented different levels of the Common European Framework of References The test was administered to 4,002 students (50% girls). Their responses were scaled using a partial credit model. Item fit statistics and differential item functioning were evaluated to ensure the quality of the test. These analyses showed that the test exhibited an acceptable reliability and a satisfactory fit to the Rasch model. Furthermore, test fairness could be confirmed for different subgroups. Limitations of the test pertained to its difficulty that that did not adequately cover the upper range of the ability distribution. Overall, the English reading test had acceptable psychometric properties that allowed for an estimation of reliable competence scores. Besides the scaling results, this paper also describes the data available in the scientific use file and presents the R syntax for scaling the data.

Keywords

item response theory, scaling, English as a foreign language, scientific use file

Content

1. Introduction

Within the National Educational Panel Study (NEPS) different competences are measured coherently across the life span. These include, among others, reading competence, mathematical competence, scientific literacy, information and communication technologies literacy, metacognition, vocabulary, and domain general cognitive functioning. An overview of the competences measured in the NEPS is given by Weinert and colleagues (2011) as well as Fuß, Gnambs, Lockl, and Attig (2016).

Most of the competence data are scaled using models based on item response theory (IRT). Because most of the competence tests were developed specifically for implementation in the NEPS, several analyses were conducted to evaluate the quality of the tests. The IRT models chosen for scaling the competence data and the analyses performed for checking the quality of the scale are described in Pohl and Carstensen (2012).

In this paper, the results of these analyses are presented for a competence test on English as a foreign language that was administered in grade 10 of starting cohort 3 (fifth grade). First, the main concepts of the English competence test are introduced. Then, the competence data of starting cohort 3 and the analyses performed on the data to estimate competence scores and to check the quality of the test are described. Finally, an overview of the data that are available for public use in the scientific use file is presented.

Please note that the analyses in this report are based on the data available at some time before public data release. Due to ongoing data protection and data cleansing issues, the data in the scientific use file (SUF) may differ slightly from the data used for the analyses in this paper. However, no fundamental changes in the presented results are expected.

2. Testing English Reading Competence

The framework and item development for the English reading competence tests was led by the Institute for Educational Quality Improvement (IQB) and is described in Rupp, Vock, Harsch, and Köller (2008). The reading competence test in English included ten short texts that were accompanied by ten item sets referring to these texts. All items were developed by trained experts and corresponded to the National Educational Standards and the Common European Framework of Reference (Council of Europe, 2001). The students had to read each text and, subsequently, answer multiple items related to this text.

The items were accompanied by different response formats (see Table 1). Simple multiple choice formats included four response options with one being correct and three response options functioning as distractors (i.e., they were incorrect). Complex multiple choice (CMC) items consisted of several subtasks that had to be rated as true, false, or information not given in the text. Matching (MA) items required the test taker to match a number of responses to a given set of statements. In all cases, there were more response options than there were statements. Examples of the different response formats are given in Pohl and Carstensen (2012) and Gehrer, Zimmermann, Artelt and Weinert (2012).

Table 1.

Number of Items by Different Response Formats

Response format	Booklet 1	Booklet 2	Booklet 3
Simple multiple choice items	0	0	4
Complex multiple choice items	2	3	2
Matching items	3	2	2
Total number of items	5	5	8

The competence test for English reading that was administered in the present study included 13 items. To evaluate the quality of these items, extensive preliminary analyses were conducted. These preliminary analyses identified a poor item fit and severe differential item functioning for one subtask in items efg10022s_sc3g10_c and efg10059s_ sc3g10_c as well as item efg10065c_sc3g10_c. Therefore, these subtasks and items were removed from the final scaling procedure. Thus, the analyses presented in the following sections and the competence scores derived for the respondents are based on the remaining 12 items.

The ten texts that accompanied the 12 items were distributed across three different booklets. Each booklet contained either five or six texts with 5 or 8 items (see Table 1). Three texts were identical in all three booklets, whereas the remaining texts were unique to each booklet (see Table 2). Each respondent was randomly assigned one of these booklets. There was no multi-matrix design regarding the order of the items *within* a specific test. All students received the test items in the same order. A detailed description of the study design, the sample, and the administered instrument is available on the NEPS website (http://www.neps-data.de).

Table 2.

Number of Items by Booklet

	Booklet 1	Booklet 2	Booklet 3
Common items	3	3	3
Unique items	2	2	5
Total number of items	5	5	8

3. Data

A total of 4,002[1] students received the English reading competence test. For 6 respondents no valid item responses were available. These cases were excluded from further analyses (see Pohl & Carstensen, 2012). Thus, the analyses presented in this paper are based on a sample of 3,996 individuals. About 50% of them were female and attended a secondary school. The students were evenly distributed among the three booklets (see Table 3).

Table 3.

Number of Participants by Booklet

	Secondary school		Other school		Total
	boys	**girls**	**boys**	**girls**	
Booklet 1	344	330	303	343	1,320
Booklet 2	364	319	296	357	1,336
Booklet 3	380	319	322	319	1,340
Total	1,088	968	921	1,019	3,922

4. Analyses

4.1 Missing Responses

Competence data include different kinds of missing responses. These are missing responses due to a) invalid responses, b) omitted items, c) items that test takers did not reach, d) items that have not been administered, and, finally, e) multiple kinds of missing responses within CMC and MA items that are not determined. Invalid responses occurred, for example, when two response options were selected although only one was required, or when numbers or letters that were not within the range of valid responses were given as a response. Omitted items occurred when test takers skipped some items. Due to time limits, not all persons finished the test within the given time. All missing responses after the last valid response given were coded as not reached. Because of the booklet design, the items unique to each booklet were not administered to participants receiving another booklet. These items were missing by design. Because CMC and MA items were aggregated from several subtasks, different kinds of missing responses or a mixture of valid and missing responses might be found in these items. A CMC or MA item was coded as missing if at least one subtask

1 Note that these numbers may differ from those found in the SUF. This is due to still ongoing data protection and data cleaning issues.

contained a missing response. If just one kind of missing response occurred, the item was coded according to the corresponding missing response. If the subtasks contained different kinds of missing responses, the item was labeled as a not determinable missing response.

Missing responses provide information on how well the test worked (e.g., time limits, understanding of instructions, handling of different response formats). They also need to be accounted for in the estimation of item and person parameters. Therefore, the occurrence of missing responses in the test was evaluated to get an impression of how well the persons were coping with the test. Missing responses per item were examined in order to evaluate how well each of the items functioned.

4.2 Scaling Model

Item and person parameters were estimated using a partial credit model (PCM; Masters, 1982). A detailed description of the scaling model can be found in Pohl and Carstensen (2012).

CMC and MA items consisted of a set of subtasks that were aggregated to a polytomous variable for each item, indicating the number of correctly sovled subtasks within that item. If at least one of the subtasks contained a missing response, the CMC or MA item was scored as missing. Categories of polytomous variables with less than N = 100 responses were collapsed in order to avoid possible estimation problems. This usually occurred for the lower categories of polytomous items; in these cases, the lower categories were collapsed into one category.

English reading competences were estimated as weighted maximum likelihood estimates (WLE; Warm, 1989). To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while simple MC items were scored dichotomously as 0 for an incorrect and 1 for the correct response (see Pohl & Carstensen, 2013, for studies on the scoring of different response formats). Person parameter estimation in NEPS is described in Pohl and Carstensen (2012), while the data available in the SUF is described in section 7.

4.3 Checking the Quality of the Test

The English reading competence test was specifically constructed for administration in the NEPS. In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses.

Before aggregating the subtasks of CMC and MA items to a polytomous variable, this approach was justified by preliminary psychometric analyses. For this purpose, the subtasks were analyzed together with the MC items in a Rasch (1960) model. The fit of the subtasks was evaluated based on the weighted mean square (WMNSQ), the respective t-value, point-biserial correlations of the correct responses with the total correct score, and the item characteristic curves. Only if the subtasks exhibited a satisfactory item fit, they were used to construct polytomous variables that were included in the final scaling model.

The MC items consisted of one correct response option and three distractors (i.e., incorrect response options). The quality of the distractors within the items was examined using the

point-biserial correlation between selecting an incorrect response option and the total correct score. Negative correlations indicate good distractors, whereas correlations between .00 and .05 are considered acceptable and correlations above .05 are viewed as problematic distractors (Pohl & Carstensen, 2012).

After aggregating the subtasks to polytomous variables, the fit of the dichotomous MC and polytomous CMC and MA items to the partial credit model (Masters, 1982) was evaluated using the weighted mean square (WMNSQ) statistic, the respective *t*-value, point-biserial correlations of the correct responses with the total correct score, and the item characteristic curves (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (*t*-value > |8|) were judged as having a considerable item misfit and their performance was further investigated. Correlations of the item score with the corrected total score (equal to the corrected discrimination) greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic. Overall judgment of the fit of an item was based on all fit indicators.

The English reading competence test should measure the same construct for all students. If some items favored certain subgroups (e.g., they were easier for males than for females), measurement invariance would be violated and a comparison of competence scores between these subgroups (e.g., males and females) would be biased and, thus, unfair. For the present study, test fairness was investigated for the variables sex, school type (i.e., gymnasium vs. other school), the number of books at home (as a proxy for socioeconomic status), and migration background (see Pohl & Carstensen, 2012a, for a description of these variables). Differential item functioning (DIF) was examined using a multigroup IRT model, in which main effects of the subgroups as well as differential effects of the subgroups on item difficulty were modeled. Based on experiences with preliminary data, we considered absolute differences in estimated difficulties between the subgroups that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and noteworthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences smaller than 0.4 as negligible DIF. Minimum hypothesis tests (see Fischer, Rohm, Gnambs, & Carstensen, 2016) were used to statistically test whether the observed differences was significantly larger than 0.4 and, thus, was at least small in size. Additionally, the test fairness was examined by comparing the fit of a model including differential item functioning to a model that only included main effects and no DIF.

The English reading competence test was scaled using the PCM (Masters, 1982) because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nonetheless, Rasch-homogeneity is an assumption that might not hold for empirical data. To test the assumption of equal item discrimination parameters, a generalized partial credit model (GPCM; Muraki 1992) was also fitted to the data and compared to the PCM.

The dimensionality of the test was evaluated by examining the residuals of the PCM. Approximately zero-order correlations as indicated by Yen's (1984) Q_3 indicate unidimensionality. Because in case of locally independent items, the Q_3 statistic tends to be slightly negative, we report the corrected Q_3 that has an expected value of 0. Following

prevalent rules-of-thumb (Yen, 1993) values of aQ_3 falling below .20 indicate essential unidimensionality.

4.4 Software

The IRT models were estimated in TAM version 2.4-9 (Kiefer, Robitzsch, & Wu, 2017) in R version 3.4.1 (R Core Team, 2017) using the Gauss-Hermite quadrature method with 21 nodes.

5. Results

5.1 Missing Responses

5.1.1 Missing responses per person

Figure 1 shows the number of invalid responses per person by booklet. There was no difference in the percentage of invalid responses between the different booklets. Overall, there were hardly any invalid responses.

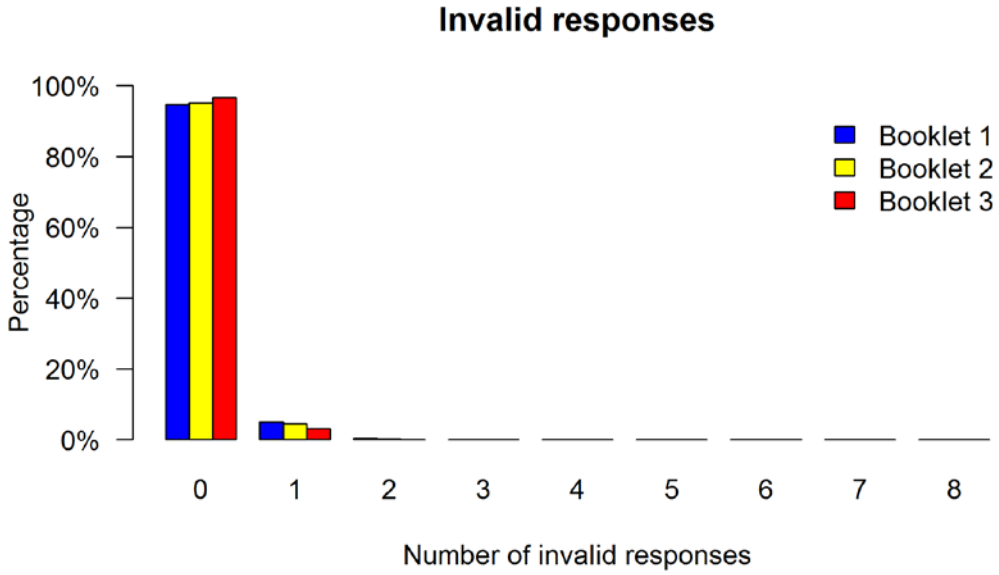

Figure 1. Number of invalid responses by booklet

Missing responses can also occur when respondents omit items. As illustrated in Figure 2 most respondents (more than 98%) did not skip any item, whereas most of the rest skipped one item.

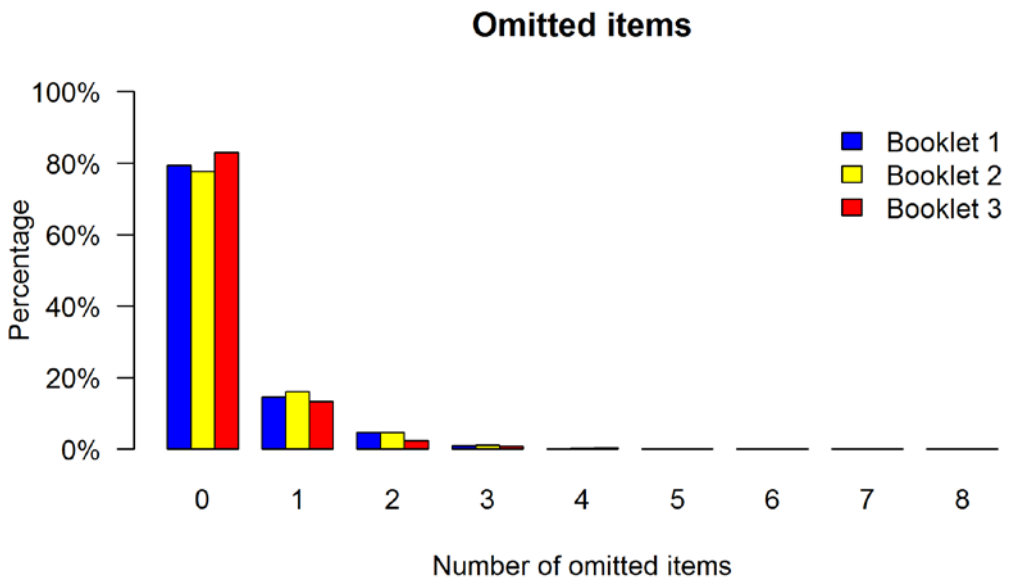

Figure 2. Number of omitted items by booklet

Another source of missing responses is items that were not reached by the respondents; these are all missing responses after the last valid response. The number of not reached items was rather low; more than 99% of the respondents finished the entire test (Figure 3). Thus, most respondents were able to finish the test within the allocated time limit.

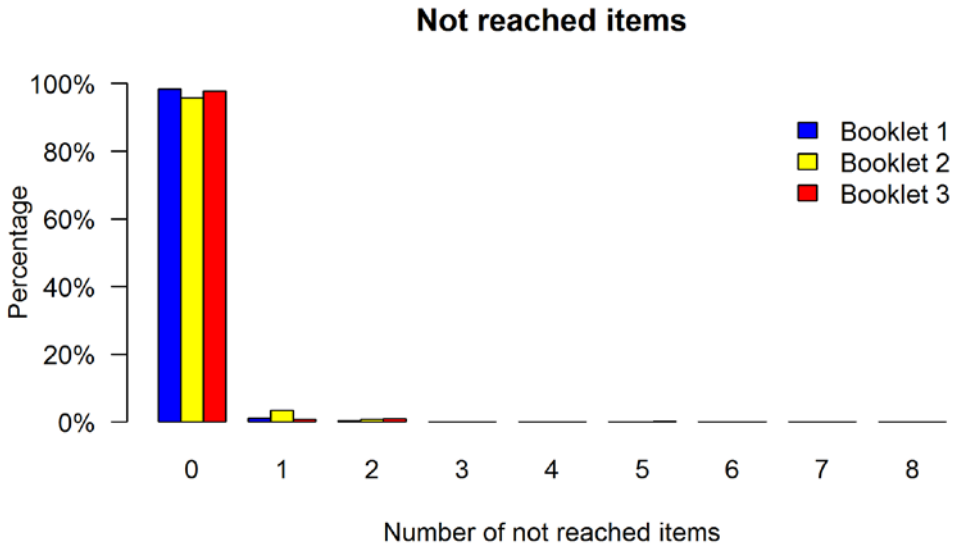

Figure 3. Number of not reached items by booklet

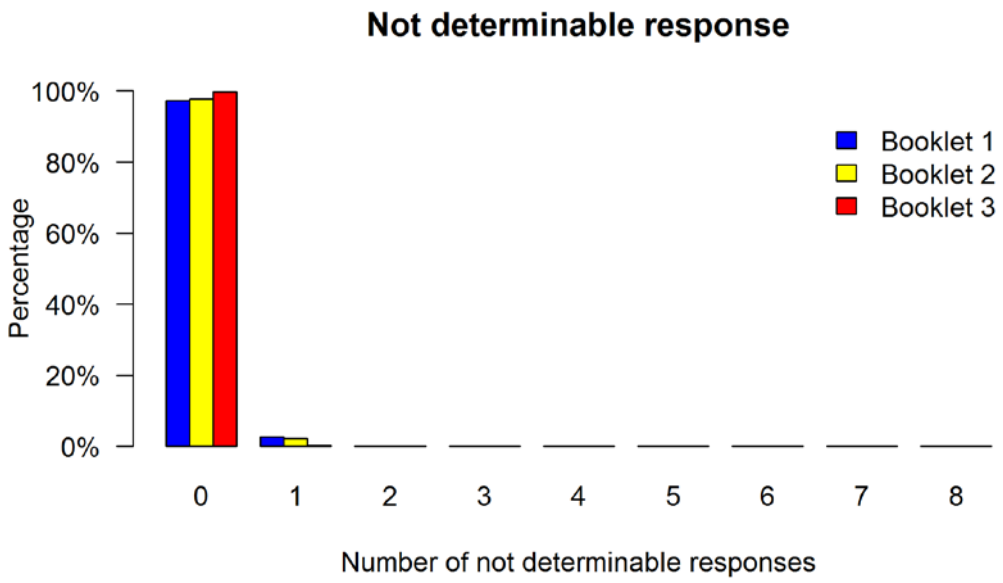

Figure 4. Number of not-determinable items by booklet

Because the CMC and MA items were aggregated from several subtasks, the missing type could not be determined for many of these items. About 20% of the respondents exhibited 1 to 2 not determinable missing values (see Figure 4).

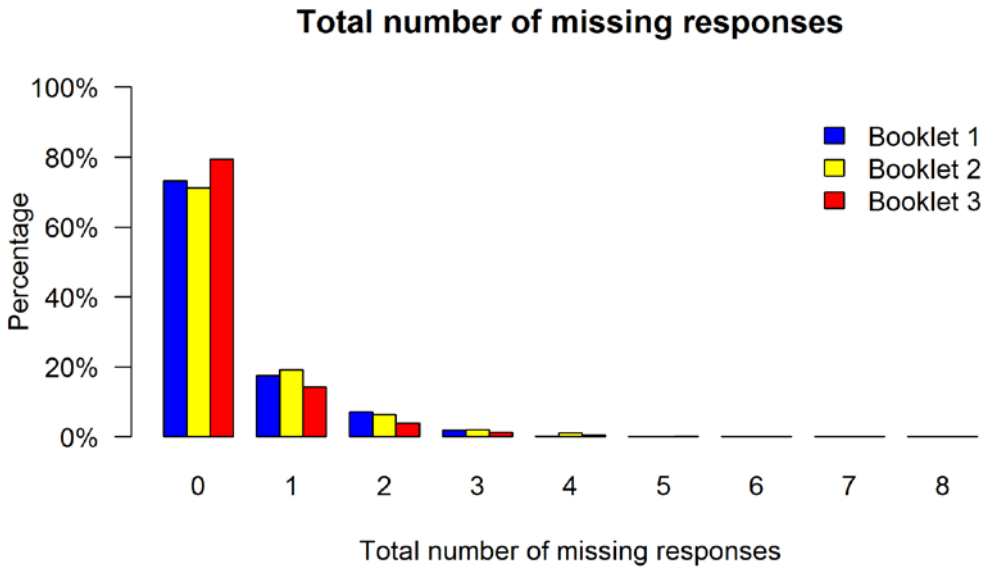

Figure 5. Total number of missing responses by booklet

The total number of missing responses, aggregated over invalid, omitted, not reached, and not determinable missing responses per person, is illustrated in Figure 5. There was no difference in the number of missing values between the three booklets. About 74% to 80% of the respondents had no missing response at all and about 6% to 8% of the participants had two or more missing responses.

In sum, the amount of missing responses was small and there were few differences in the number of missing responses between the three booklets.

5.1.2 Missing responses per item

Table 4 provides information on the occurrence of different kinds of missing responses per item for the three booklets. Overall, in the three booklets the number of not determinable responses varied across items between 0.00% and 1.97% and were, thus, negligible. In contrast, there were more omitted responses. In particular, item efg10059s_c in booklet 1 exhibited a rather large amount for missing responses (13%) as compared to the remaining items (between 0.00% and 9.06%). Thus, it seems that some students had problems with this MA item. In contrast, the percentage of invalid responses per item (columns "NV" in Table 4) was rather low, with the maximum rate being 2.73%.

Table 4.

Percentage of Missing Values by Item.

		Booklet 1					Booklet 2					Booklet 3				
	Item	N	NR	OM	NV	ND	N	NR	OM	NV	ND	N	NR	OM	NV	ND
1.	efg10022s_sc3g10_c	1,250	0.00	5.23	0.08	0.00	1,243	0.00	6.59	0.30	0.07	1,257	0.00	6.04	0.15	0.00
2.	efg10108s_sc3g10_c	1,193	0.00	7.05	2.20	0.38	1,184	0.00	9.06	2.10	0.22	1,208	0.00	8.06	1.79	0.00
3.	efg10094s_sc3g10_c	1,287	0.08	2.12	0.23	0.08	1,296	0.00	2.69	0.30	0.00	1,305	0.07	2.39	0.15	0.00
4.	efg10059s_sc3g10_c	1,079	0.45	13.11	2.73	1.97										
5.	efg10002s_sc3g10_c	1,284	1.59	0.38	0.53	0.23										
6.	efg10008s_sc3g10_c						1,167	0.82	8.46	2.25	1.12					
7.	efg10098s_sc3g10_c						1,219	4.27	3.52	0.15	0.82					
8.	efg10065a_sc3g10_c											1,313	0.30	1.49	0.22	0.00
9.	efg10065b_sc3g10_c											1,314	0.30	1.64	0.00	0.00
10.	efg10065d_sc3g10_c											1,314	0.37	1.49	0.07	0.00
11.	efg10075s_sc3g10_c											1,286	1.42	1.64	0.75	0.22
12.	efg10057a_sc3g10_c											1,305	2.24	0.00	0.37	0.00

Note. N = Number of valid responses, NR = Percentage of respondents that did not reach item, OM = Percentage of respondents that omitted the item, NV = Percentage of respondents with an invalid response, ND = Percentage of respondents with a not-determinable response.

With an item's progressing position in the test, the amount of persons that did not reach an item (columns "NR" in Table 4) rose to about 4% in booklet 2. However, in all booklets, the last items were not reached by some respondents (see Figure 6). Overall, the percentage of respondents that did not reach an item was rather low.

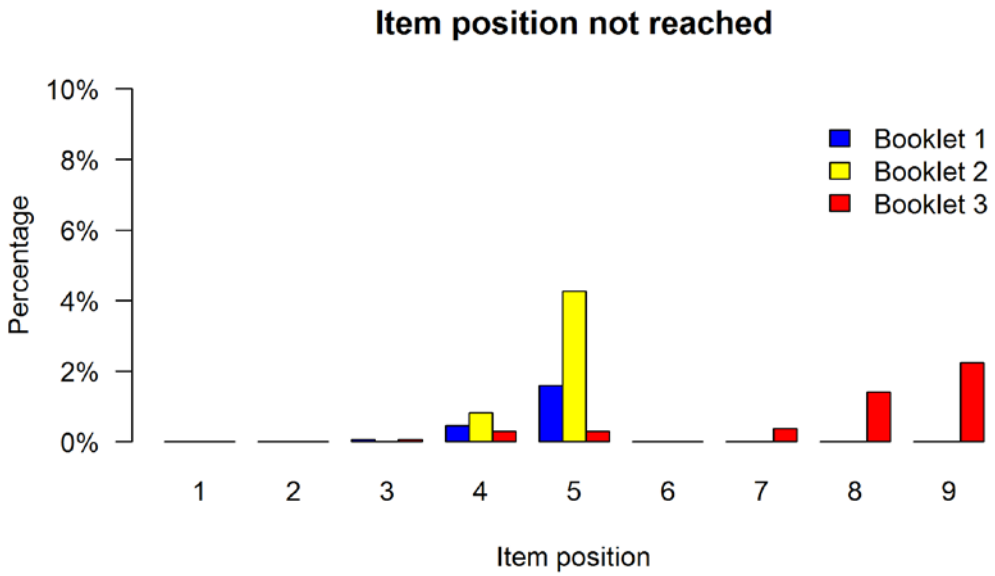

Figure 6. Item position not reached by booklet (booklets 1 and 2 included five items, whereas booklet 3 included nine items).

5.2 Parameter Estimates

5.2.1 Item parameters

The third column in Table 5 presents the percentage of correct responses (for simple multiple choice items) in relation to all valid responses for each item. Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index of item difficulty. The percentage of correct responses varied between 53% and 65% with an average of 60% (*SD* = 6%) correct responses.

The estimated item difficulties (for dichotomous variables) and location parameters (for polytomous variables) are given in Table 5. The step parameters for polytomous variables are summarized in Table 6. The item difficulties and location parameters were estimated by constraining the mean of the ability distribution to be zero. Due to the large sample size, the standard errors (SE) of the estimated parameters (see Tables 5 and 6) were rather small (all SEs ≤ 0.10). The estimated item difficulties and location parameters ranged from -1.2 (item efg10002s_sc3g10_c) to 0.1 (item efg10008s_sc3g10_c). Thus, they covered a rather limited range; particularly, there were no items with high difficulty or location parameters.

Table 5.

Item Parameters

	Item	Position	Percentage correct	Difficulty	SE	WMNSQ	t	r_{it}	Discr.	Q_3
1.	efg10022s_sc3g10_c	1		-0.35	0.01	1.02	1.06	.44	0.67	.05
2.	efg10108s_sc3g10_c	2		-0.82	0.02	0.97	-1.12	.43	0.82	.04
3.	efg10094s_sc3g10_c	3		-0.74	0.01	0.97	-1.43	.43	0.80	.06
4.	efg10059s_sc3g10_c	4		0.05	0.02	0.97	-0.59	.61	0.87	.10
5.	efg10002s_sc3g10_c	5		-1.20	0.05	0.86	-3.15	.44	1.32	.08
6.	efg10008s_sc3g10_c	4		0.10	0.03	0.95	-1.28	.56	0.80	.03
7.	efg10098s_sc3g10_c	5		0.01	0.02	0.98	-0.61	.58	0.82	.07
8.	efg10065a_sc3g10_c	4	54	-0.26	0.06	1.07	2.61	.36	1.13	.04
9.	efg10065b_sc3g10_c	5	64	-0.83	0.07	0.94	-2.22	.46	1.79	.06
10.	efg10065d_sc3g10_c	7	53	-0.18	0.06	1.07	2.59	.36	1.01	.03
11.	efg10075s_sc3g10_c	8		0.09	0.03	1.11	3.21	.36	0.47	.02
12.	efg10057a_sc3g10_c	9	65	-0.83	0.07	1.13	4.13	.30	0.84	.02

Note. Difficulty = Item difficulty / location, SE = Standard error of item difficulty / location, WMNSQ = Weighted mean square, t = t-value for WMNSQ, r_{it} = Corrected item-total correlation, Discr. = Discrimination parameter of a generalized partial credit model, Q_3 =Average absolute residual correlation for item (Yen, 1983).

The item on position 6 in booklet 3 was excluded from the analyses due to an unsatisfactory item fit (see section 2). Percent correct scores are not informative for polytomous CMC and MC item scores and, therefore, are not reported.

5.2.2 Test targeting and reliability

Test targeting focuses on comparing the item difficulties with the person abilities (WLEs) to evaluate the appropriateness of the test for the specific target population. Because most items in the English reading test were polytomous, we calculated Thurstonian thresholds for each response category (Wu, Adams, Wilson, & Haldane, 2007). These indicate the location at the latent dimension at which the probability of achieving a score above the respective threshold is 50%. Thus, it is similar to the item difficulties of dichotomous items. In Figure 6, the category thresholds of the English reading items and the ability of the test takers are plotted on the same scale. The distribution of the estimated test takers' ability is mapped onto the left side whereas the right side shows the distribution of category thresholds. The respective thresholds ranged from -4.53 (item efg10108s_sc3g10_c) to 3.31 (item efg10098s_sc3g10_c) and, thus, spanned a rather broad range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 2.09, which

implies good differentiation between subjects. The reliability of the test (EAP/PV reliability = .75, WLE reliability = .68) was acceptable. The mean of the category threshold distribution was about 0.58 logits below the mean person ability distribution. Thus, although the items covered a wide range of the ability distribution, the items were slightly too easy. As a consequence, person ability in medium- and low-ability regions will be measured relative precisely, whereas higher ability estimates will have larger standard errors of measurement.

Table 6

Step Parameters (with Standard Errors) for Polytomous Items

Item	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
efg10022s_sc3g10_c	-1.06 (0.04)	-0.68 (0.04)	-0.12 (0.04)	0.09 (0.04)	0.54 (0.04)	1.23
efg10108s_sc3g10_c	-1.02 (0.04)	-0.66 (0.03)	0.56 (0.04)	1.12		
efg10094s_sc3g10_c	-0.56 (0.04)	-0.63 (0.04)	-0.32 (0.03)	-0.14 (0.03)	0.45 (0.04)	1.20
efg10059s_sc3g10_c	-0.85 (0.07)	-0.18 (0.07)	-0.21 (0.07)	0.08 (0.08)	0.26 (0.10)	0.89
efg10002s_sc3g10_c	0.32 (0.08)	-0.32				
efg10008s_sc3g10_c	-0.71 (0.06)	0.30 (0.07)	0.62 (0.10)	-0.21		
efg10098s_c3g10_c	-0.78 (0.07)	-0.62 (0.07)	-0.40 (0.06)	0.08 (0.07)	0.38 (0.08)	1.34
efg10075s_sc3g10_c	0.20 (0.06)	-0.36 (0.07)	0.16			

Note. The last step parameter for each item is not estimated and has, thus, no standard error because it is a constrained parameter for model identification.

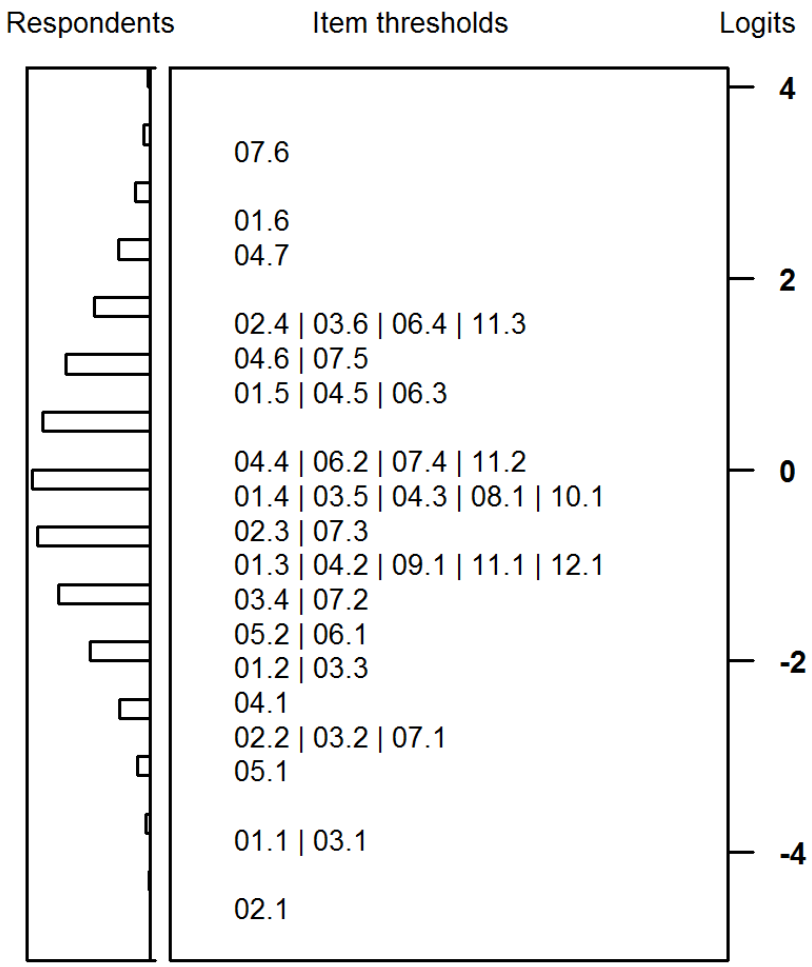

Figure 6. Test targeting. The distribution of person ability in the sample is given on the left-hand side of the graph. The category thresholds of the items are given on the right-hand side of the graph. Each number represents one threshold with the first part (before the dot) corresponding to the item number in Table 5 and the second part indicating the threshold.

5.3 Quality of the test

5.3.1 Item fit

The evaluation of the item fit was performed based on the final scaling model, the PCM. Altogether, item fit was good (see Table 5). Values of the WMNSQ ranged from 0.86 (item efg10002s_sc3g10_c) to 1.13 (item efg10057a_sc3g10_c). No item exhibited a *t*-value of the WMNSQ greater than 6. Moreover, a visual inspection of the item characteristic curves did not indicate severe abnormalities for any item. Point-biserial correlations between the item scores and the total rest scores ranged from .30 (item efg10057a_sc3g10_c) to .61 (item efg10059s_sc3g10_c) and had a mean of .45.

5.3.2 Distractor analyses

In addition to the overall item fit, it was investigated how well the distractors performed in the test by evaluating the point-biserial correlation between each incorrect response

(distractor) and the students' total correct score. The point-biserial correlations for the distractors ranged from -.41 to .04 with a mean of -.17. These results indicate that the distractors functioned well.

5.3.3 Differential item functioning

Differential item functioning (DIF) was used to evaluate test fairness for several subgroups (i.e., measurement invariance). For this purpose, DIF was examined for the variables gender, the number of books at home (as a proxy for socioeconomic status), migration background, and school type (see Pohl & Carstensen, 2012, for a description of these variables). In addition, the measurement invariance of the common items that were administered to all participants were examined across the three booklets. The differences between the estimated item difficulties in the various groups are summarized in Table 7. For example, the column "Male vs. female" reports the differences in item difficulties between men and women; a positive value would indicate that the test was more difficult for males, whereas a negative value would highlight a lower difficulty for males as opposed to females. Besides investigating DIF for each single item, an overall test for DIF was performed by comparing models which allow for DIF to those that only estimate main effects (see Table 8).

Gender: The sample included 2,009 boys and 1,987 girls. On average, female participants had a slightly higher estimated English reading ability than males (main effect = 0.14 logits, Cohen's d = 0.17). No item showed DIF greater than 0.4 logits. An overall test for DIF (see Table 8) was conducted by comparing the DIF model to a model that only estimated main effects (but ignored potential DIF). A model comparison using Akaike's (1974) information criterion (AIC) favored the more parsimonious model including only the main effect. Similar results were obtained using the Bayesian information criterion (BIC; Schwarz, 1978) that takes the number of estimated parameters into account and, thus, guards against overparameterization of models. Thus, overall, there was no pronounced DIF with regard to sex.

Books: The number of books at home was used as a proxy for socioeconomic status. There were 1,658 test takers with 0 to 100 books at home and 2,320 test takers with more than 100 books at home. There were considerable average differences between the two groups. Participants with 100 or less books at home performed on average 0.51 logits (Cohen's d = 0.67) lower in reading than participants with more than 100 books. There was no considerable DIF comparing participants with many or fewer books (highest DIF = 0.37 for item efg10098s_sc3g10_c). As a consequence, also the overall test for DIF using the BIC favored the main effects model (Table 8).

Table 7

Differential Item Functioning

Item	Gender	Books	Migration	School	Booklet		
	male vs. female	< 100 vs. ≥ 100	without vs. with	no sec. vs. sec.	1 vs. 2	1. vs. 3.	2 vs.3.
efg10022s_sc3g10_c	-0.13 (-0.16)	0.02 (0.03)	0.03 (0.04)	0.24 (0.36)	-0.03 (-0.04)	0.01 (0.02)	0.04 (0.05)
efg10108s_sc3g10_c	-0.02 (-0.03)	-0.04 (-0.06)	0.26 (0.32)	-0.14 (-0.21)	0.04 (0.05)	0.04 (0.05)	-0.01 (-0.01)
efg10094s_sc3g10_c	-0.07 (-0.08)	-0.10 (-0.14)	0.06 (0.07)	-0.11 (-0.16)	-0.01 (-0.02)	-0.05 (-0.07)	-0.04 (-0.05)
efg10059s_sc3g10_c	0.04 (0.05)	0.19 (0.25)	0.17 (0.22)	-0.04 (-0.05)			
efg10002s_sc3g10_c	-0.11 (-0.13)	-0.15 (-0.20)	0.08 (0.10)	-0.02 (-0.03)			
efg10008s_sc3g10_c	0.16 (0.20)	0.01 (0.02)	-0.29 (-0.36)	0.01 (0.02)			
efg10098s_sc3g10_c	0.28 (0.35)	0.37 (0.49*)	-0.20 (-0.26)	0.69 (1.03*)			
efg10065a_sc3g10_c	0.06 (0.07)	-0.06 (-0.07)	0.25 (0.31)	-0.02 (-0.03)			
efg10065b_sc3g10_c	0.08 (0.10)	-0.16 (-0.21)	-0.05 (-0.07)	-0.23 (-0.34)			
efg10065d_sc3g10_c	-0.09 (-0.11)	0.01 (0.01)	-0.17 (-0.22)	-0.12 (-0.18)			
efg10075s_sc3g10_c	-0.14 (-0.17)	-0.01 (-0.01)	0.00 (0.00)	-0.17 (-0.26)			
efg10057a_sc3g10_c	-0.05 (-0.07)	-0.08 (-0.11)	-0.14 (-0.17)	-0.10 (-0.14)			
Main effect (DIF model)	-0.14 (-0.17)	-0.51 (-0.67)	0.23 (0.29)	-0.94 (-1.40)	-0.03 (-0.04)	-0.01 (-0.02)	0.02 (0.02)
Main effect (Main effect model)	-0.08 (-0.10)	-0.46 (-0.60)	0.25 (0.32)	-0.85 (-1.26)	-0.02 (-0.03)	-0.01 (-0.01)	0.02 (0.02)

Note. Raw differences between item difficulties with standardized differences (Cohen's *d*) in parentheses. Sec. = Secondary school (German: „Gymnasium").

* Absolute standardized difference is significantly, $p < .05$, greater than 0.4 (see Fischer et al., 2016).

Table 8

Comparisons of Models with and without DIF

DIF variable	Model	*N*	Deviance	Number of parameters	AIC	BIC
Gender	DIF	3,996	58,281	58	58,397	58,762
	main effect	3,996	58,301	47	58,395	58,691
Books	DIF	3,978	57,780	58	57,496	58,260
	main effect	3,978	57,812	47	57,515	58,202
Migration	DIF	3,932	57,380	58	57,896	57,860
	main effect	3,932	57,421	47	57,906	57,810
School type	DIF	3,996	57,272	58	57,389	57,754
	main effect	3,996	57,330	47	57,424	57,720
Common items	DIF	3,996	35,686	21	35,736	35,893
	main effect	3,996	35,690	25	35,732	35,864

Migration background: There were 3,422 participants with no migration background and 510 subjects with a migration background. In comparison to subjects without migration background, participants with migration background had, on average, a slightly lower English reading ability (main effect = 0.23 logits, Cohen's d = 0.29). There was no noteworthy item DIF due to migration background; differences in estimated difficulties did not exceed 0.4 logits (highest DIF = -0.29 for item efg10008s_sc3g10_c). Moreover, the overall test for DIF using the BIC also favored the main effects model that did not include item-level DIF.

School type: Overall, 1,940 subjects who took the reading test attended secondary school (German: "Gymnasium") whereas 2,056 were enrolled in other school types. Subjects in secondary schools showed a higher reading ability in English on average (0.94 logits, Cohen's d = 1.40) than subjects in other school types. One item (efg10098s_sc3g10_c) exhibited noteworthy item DIF (DIF = 0.69 logits). However, the DIF was not considered severe. Moreover, the overall model test using the BIC indicated a slightly better fit for the more parsimonious main effects model that did not account for item level DIF.

Common items: The participants received different booklets with different tests (see Table 3). Only a subset of three items that were included in all three booklets was administered to all participants. For these common items, potential DIF was examined between the three booklets. As expected, there were no pronounced differences in the subjects' mean abilities between the three booklets (Cohen's d between -0.02 and 0.04). There was no noteworthy DIF for the common items with regard to the three booklets. Also, the overall tests for DIF favored the main effects model that did not include item-level DIF.

5.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1980) model is that all item-discrimination parameters are equal. In order to test this assumption, a generalized partial credit model (GPCM; Muraki, 1992) that estimates discrimination parameters was fitted to the data. The estimated discrimination parameters differed moderately among items (see Table 5). The average discrimination parameter fell at 0.97. Particularly, the discrimination parameter of 0.47 for item efg10075s_sc3g10_c was rather low. However, an inspection of the respective item characteristic curve of the PCM indicated an adequate fit. Model fit indices suggested a slightly better model fit of the GPCM (AIC = 58,231, BIC = 58,571, number of parameters = 58) as compared to the PCM (AIC = 58,373, BIC = 58,643, number of parameters = 43). Despite the empirical preference for the GPCM, the PCM more adequately matches the theoretical conceptions underlying the test construction (see Pohl & Carstensen, 2012, 2013, for a discussion of this issue). For this reason, the PCM was chosen as our scaling model to preserve the item weightings as intended in the theoretical framework.

5.3.5 Unidimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the PCM. The adjusted Q_3 statistics (see Table 5) were quite low (M = .05, SD = .02)—the largest individual residual correlation was .10—and, thus, indicated an essentially unidimensional test. Because the reading test is constructed to measure a single dimension, a unidimensional reading competence score was estimated.

6. Discussion

The analyses in the previous sections reported information on the quality of the English reading test in starting cohort 3 for grade 10 and described how the reading competence scores were estimated. Different kinds of missing responses were examined, item fit statistics were thoroughly checked, and the correlations between the responses and the total correct scores were investigated. Further quality inspections were conducted by examining differential item functioning and testing Rasch-homogeneity. Various criteria indicated a good fit of the items and measurement invariance across various subgroups. Moreover, the number of missing responses were reasonably small. The test had a high reliability and distinguished well between test takers. However, the test was better targeted at mediocre- and low-performing students and did not sufficiently cover the high ability spectrum. As a consequence, ability estimates will be precise for low-performing students but less precise for high performing students. In summary, the test had good psychometric properties that allowed the estimation of a unidimensional reading competence score for English as a foreign language.

7. Data in the Scientific Use File

7.1 Naming conventions

The SUF contains 13 items, of which 5 were scored dichotomously (MC items) with 0 indicating an incorrect response and 1 indicating a correct response. These items are marked with a ‘0_c’ at the end of the variable name. A total of 8 items were scored as polytomous variables (CMC and MA items) that are marked with a ‘s_c’ at the end of the variable names.

For further details on the naming conventions of the variables see Fuß and colleagues (2016).

7.2 Linking of competence scores

In grade 10 of starting cohort 4 the identical English competence test has been administered. However, the competence scores derived in the different starting cohorts cannot be directly compared, because differences in observed scores would reflect differences in competences as well as differences in test difficulties. To place the different measurements onto a common scale and, thus, allow for meaningful comparisons across cohorts, the linking procedure described in Fischer, Rohm, Gnambs, and Carstensen (2016) was adopted. Following an anchor-items design, the responses from starting cohort 3 were linked to the scale of test administered in starting cohort 4.

7.2.1 Samples

In starting cohort 3, 3,996 individuals provided valid responses to the English reading competence test, whereas 10,868 respondents were available in starting cohort 4. These respondents were used to link the two tests across both starting cohorts (see Fischer et al., 2016.).

7.2.2 Differential Item Functioning

Items that are supposed to link two tests must exhibit measurement invariance; otherwise, they cannot be used for the linking procedure. Therefore, we tested whether the item parameters derived in the two samples showed a non-negligible shift in item difficulties. The differences in item difficulties between starting cohorts 3 and 4 and the tests for measurement invariance based on the Wald statistic (see Fischer et al., 2016) are summarized in Table 9. For 10 items (difference in logits: *Min* = -0.08, *Max* = 0.16) measurement invariance was supported, that is, the minimum effects hypothesis test was not significant (α = .05). However, items efg10022s_sc3g10_c and efg10065a_sc3g10_c were significantly easier in starting cohort 3 (difference in logits: -0.23 and -0.24). Therefore, these two items were not used for linking the two starting cohorts. The English reading competence tests administered in the two starting cohorts were linked using the "mean/mean" method for the anchor-items design using the 10 items without DIF (see Fischer et al., 2016).

The correction term was calculated as $c = -0.201$. This correction term was subsequently added to each difficulty parameter estimated in starting cohort 3 (see Table 5) to derive the linked item parameters. The link error reflecting the uncertainty in the linking process was calculated according to equation 2 in Fischer et al. (2016) as 0.049 and has to be included into the *SE* when statistical tests are used to compare groups concerning their mean change of ability between two linked measurements.

Table 9

Differential Item Functioning Analyses between Starting Cohorts 3 and 4.

	Item	Δσ	$SE_{\Delta\sigma}$	*F*
1.	efg10022s_sc3g10_c	-0.23	0.01	326.58*
2.	efg10108s_sc3g10_c	-0.08	0.01	48.78
3.	efg10094s_sc3g10_c	0.06	0.01	62.29
4.	efg10059s_sc3g10_c	0.16	0.01	148.40
5.	efg10002s_sc3g10_c	-0.02	0.01	2.61
6.	efg10008s_sc3g10_c	0.12	0.01	77.14
7.	efg10098s_sc3g10_c	0.04	0.01	8.85
8.	efg10065a_sc3g10_c	-0.24	0.01	328.33*
9.	efg10065b_sc3g10_c	0.10	0.01	13.80
10.	efg10065d_sc3g10_c	0.02	0.01	8.06
11.	efg10075s_sc3g10_c	0.04	0.01	3.31
12.	efg10057a_sc3g10_c	0.11	0.04	3.07

Note. Δσ = Difference in item difficulty parameters between the starting cohorts (negative values indicate easier items in starting cohort 3); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 14,880) = 285.64. A non-significant test indicates measurement invariance.

*$p < .05$

7.3 English reading competence scores

In the SUF, manifest English reading competence scores are provided in the form of WLEs ("efg10_sc1") including their respective standard error ("efg10_sc2"). The R Syntax for estimating the WLEs is provided in Appendix A. In the IRT scaling model, the polytomous CMC and MA variables were scored as 0.5 for each category. In case less than 50% of subtasks in a PCM item were missing, these missing values were imputed with the expected score from the Rasch analyses presented above. Subsequently, the PCM scores were recalculated based on the imputed values. No imputations were performed if more than 50% of subtasks were missing for a given respondent. For students who did not take part in the reading test or who did not give enough valid responses no WLEs were estimated. The value on the WLE and the respective standard error for these persons are denoted as not-determinable missing values. Alternatively, users interested in examining latent relationships may either include the measurement model in their analyses or estimate plausible values. A description of these approaches can be found in Pohl and Carstensen (2012).

References

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, United Kingdom: University Press.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). *Linking the data of the competence tests* (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Fuß, D., Gnambs, T., Lockl, K., & Attig, M. (2016). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *The assessment of reading competence (including sample items for grade 5 and 9)*. Scientific Use File 2012, Version 1.0.0. Bamberg: University of Bamberg, National Educational Panel Study.

Haberkorn, K., Pohl, S., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 4 in Ninth Grade* (NEPS Working Paper No. 16). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Hardt, K., Pohl, S., Haberkorn, K., & Wiegand, E. (2013). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 6 for adults in main study 2010/11* (NEPS Working Paper No. 25). Bamberg: University of Bamberg, National Educational Panel Study.

Kiefer, T., Robitzsch, A. & Wu, M. (2017). *TAM: Test analysis modules*. R package version 1.99999-31. URL: https://CRAN.R-project.org/package=TAM

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. doi:10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. ETS *Applied Psychological Measurement, 16*, 159-176. doi:10.1177/014662169201600206

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests.* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189-216.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Rupp, A. A., Vock, M., Harsch, C., & Köller, O. (2008*). Developing standards-based assessment tasks for English as a first foreign language: context, processes, and outcomes in Germany* (Vol. 1). Waxmann.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450. doi:10.1007/BF02294627

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14*, 67-86. doi:10.1007/s11618-011-0182-7

Wu, Adams, Wilson, & Haldane, 2007. *ACER ConQuest Version 2.0*. Camberwell, Australia: Acer Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Appendix

Appendix A: R-Syntax for estimating WLEs in grade 10 of starting cohort 3

```
# load packages
library(haven) # to import SPSS files
library(TAM)   # for IRT analyses

# load competence data
dat <- read_sav("SUF for competencies in SC 3.sav")

# items of the English competence test
items <- c("efg10022s_sc3g10_c", "efg10108s_sc3g10_c",
           "efg10094s_sc3g10_c", "efg10059s_sc3g10_c",
           "efg10002s_sc3g10_c", "efg10008s_sc3g10_c",
           "efg10098s_sc3g10_c", "efg10065a_sc3g10_c",
           "efg10065b_sc3g10_c", "efg10065d_sc3g10_c",
           "efg10075s_sc3g10_c", "efg10057a_sc3g10_c")

# define Q-matrix for 0.5 scoring of PCM
Q <- matrix(1, nrow = length(items), ncol = 1)
Q[c(1:7, 11), 1] <- 0.5    # score of 0.5 for polytomous items

# estimate partial credit model
mod <- tam.mml(resp = dat[, items], Q = Q, irtmodel = "PCM2",
               pid = dat$ID_t)
summary(mod)

# item fit
tam.fit(mod)

# WLE
tam.wle(mod)
```